

Equivalence classes of random Boolean trees^{*}

Antoine Genitrini¹ and Cécile Mailler²

¹ Laboratoire d'Informatique de Paris 6; Antoine.Genitrini@lip6.fr.

² Laboratoire de Mathématiques de Versailles; Cecile.Mailler@uvsq.fr.

Abstract. An *and/or* tree is a binary plane tree, with internal nodes labelled by logical connectives, and with leaves labelled by literals chosen in a fixed set of k variables and their negations. Pick up uniformly at random such a Boolean tree with n leaves, and consider the Boolean function it represents. Finally, let the size n of the trees tend to infinity. This process defines a random distribution on Boolean functions of k variables, named the Catalan tree distribution. It has long been studied in the literature, however quantitative results were obtained only by taking in a last step an infinite limit for k .

In the present paper, we investigate the global model such that the number of variables k_n is a function of n . We describe the whole range of the probability distributions depending on the function k_n , as soon as it tends jointly with n to infinity. In this context, we exhibit a threshold M_n , equivalent to $n/\ln n$, such that, when k_n becomes larger, then the probability distribution becomes stable.

To study this model, we mainly use analytic combinatorics and we extend the Kozik's *pattern theory*, first developed for the Catalan tree model.

Keywords: Random Boolean expressions; Boolean formulas; Boolean functions; Probability distribution; Analytic combinatorics; Complexity.

1 Introduction

Pick up uniformly at random a large Boolean expression and focus on the Boolean function it represents. How random is this Boolean function? E.g., what is the probability to get a satisfiable function? or any given function? Former results based on specific Boolean expressions (the variables and the connectives used to build the expressions are fixed and finite sets) highlight a relation between the *complexity* of a function and its probability.

The first approach, by Lefmann and Savický [1], consists in fixing a finite set of variables, allowing the two logical connectives *and* and *or* and choosing uniformly at random a Boolean expression of *size* n in this logical system. Lefmann and Savický first proved the existence of a limiting probability distribution on Boolean functions when the size of the random Boolean expression tends to infinity. Since the seminal paper by Chauvin et al. [2], almost all quantitative studies of such a Boolean distributions are deeply related to analytic combinatorics: a survey by Gardy [3] provides a wide range of models with various

^{*} Partially supported by the A.N.R. project *BOOLE*, 09BLAN0011.

numerical results. Later, Kozik [4] proved a strong relation between the limiting probability of a given function and its *complexity* (i.e. the minimal *size* of an expression representing the function). His approach lies in two steps: (1) first let the size of the Boolean expressions taken into consideration tend to infinity, and then (2) let the number of variables used to label the expressions tend to infinity. His powerful machinery, the *pattern theory*, easily classifies and counts large expressions according to structural constraints. This theory, already adapted to other logical systems [5]³ will also be extended in this paper.

In order to *swap* the two ordered limits (1) and (2) (on the size of the expressions and then on the number of variables) Genitrini *et al.* [6,7] presented another model of random expressions built on a infinite set of variables: a notion of equivalence classes of expressions is needed and introduced by the authors. Though some interesting similarities between this new model and the *finite* one have been observed, no direct link has been explained.

This paper presents a more general model, which unifies in a unique approach both previous models. By using a slightly different equivalence relationship on Boolean expressions, we manage to let both the number of variables and the size of the formulas, tend jointly to infinity. We let the number of variables be a function of the size of the expressions and exhibit some threshold: as soon as the number of variables is *large enough* compared to the size of the expressions, the general behaviour of the induced probability on the set of Boolean functions does not change anymore by adding more variables.

We focus on the logical context of **and/or** connectives in order to adapt the pattern theory of Kozik and because of the richness of this logical system (normal forms, functional completeness). However the implicational logical system (e.g. [8,7]) could also be studied in this new context and we deeply believe the general behaviour to be identical.

The paper is organized as follows. Section 2 introduces our unified model based on equivalence relation of Boolean expressions. Then, Section 3 states our two main results: (1) the link between the probability of a class of functions and the complexity of the functions taken into account; (2) the behaviour of the probability related to the dynamic between the number of variables and the size of the expressions. Section 4 is devoted to the technical core of the paper. Finally Section 5 applies our approach to **and/or** trees and proves the main results.

Almost all proofs are given in the appendices.

2 Probability distributions on equivalence classes of Boolean functions

2.1 Contextual definitions

A Boolean function is a function from $\{0,1\}^{\mathbb{N}}$ into $\{0,1\}$. The set of Boolean functions is denoted by \mathcal{F} . In the following, $\{x_1, x_2, \dots\}$ will be an element of

³ Genitrini *et al.* adapted the pattern language theory to associative or commutative connectives: cf. preprint at <http://lip6.fr/Antoine.Genitrini/GGKM.pdf>.

$\{0, 1\}^{\mathbb{N}}$. A variable x_i can be negated: $\bar{x}_i = 1 - x_i$, and we call **literal** a variable or its negation. The two connectives taken into account, **and** and **or**, are respectively denoted by \wedge and \vee .

An **and/or** Boolean expression is seen as an **and/or tree** i.e. a binary plane tree with leaves labelled by a literal and with internal nodes labelled by a connective. Each **and/or tree** computes (or represents) a Boolean function. Obviously an infinite number of **and/or trees** are computing the same Boolean function. The **size** of an **and/or tree** is its number of leaves: remark that, for all $n \geq 1$, there is an infinite number of **and/or trees** of size n .

The **complexity** of a Boolean function f , denoted by $L(f)$, is defined as the size of its **minimal trees**, i.e. the smallest trees computing f . Although a Boolean function is defined on an infinite set of variables, it may *really* depend only on a finite subset of *essential variables*: given a Boolean function f , we say that the variable x is **essential** for f , if and only if $f|_{x \leftarrow 0} \neq f|_{x \leftarrow 1}$ (where $f|_{x \leftarrow \alpha}$ is the restriction of f to the subspace of $\{0, 1\}^{\mathbb{N}}$ where $x = \alpha$). We denote by $E(f)$ the number of essential variables of f . Remark that the complexity and the number of essential variables of a Boolean function are only related by the following inequality: $E(f) \leq L(f)$.

2.2 Equivalence relations

Analytic combinatorics' tools (cf. [9]) are based on the notion of combinatorial classes. A *combinatorial class* is a denumerable (or finite) set of objects on which a size notion is defined such that each object has a non-negative size and the set of objects of any given size is finite. Thus our class of **and/or trees** is not a combinatorial class. To use the analytic combinatorics' tools, we define an equivalence probability distribution on Boolean trees.

The following equivalence relation is distinct from the one of [6,7], because their logical context does not allow negated variables. In the rest of the paper, we define a **tree-structure** to be an **and/or tree** in which leaves labels have been removed (but internal nodes remain labelled).

Definition 1. *Let A and B be two **and/or trees**. Trees A and B are **equivalent** if (1) their tree-structures are identical, (2) two leaves are labelled by the same variable in A if and only if they are labelled by a same variable in B , and (3) two leaves are labelled by the same literal in A if and only if they are labelled by a same literal in B .*

This equivalence relationship on Boolean trees *induces straightforwardly an equivalence relationship on Boolean functions*.

For example, both functions $(x_i)_{i \geq 1} \mapsto \bar{x}_{2013}$ and $(x_i)_{i \geq 1} \mapsto x_1$ are equivalent. An important remark is that all functions of an equivalence class have the same complexity and the same number of essential variables. In the following, we will denote by $\langle f \rangle$ the equivalence class of the Boolean function f .

2.3 Probability distribution

In the following, k_n is the maximum number of different variables that can appear as labels of a **and/or** tree of size n . We assume that the sequence \mathbf{k}_n is **increasing and tends to infinity** as n tends to infinity.

Definition 2. We denote by T_n the number of equivalence classes of trees of size n in which at most k_n different variables appear as leaves labels. We define the ordinary generating function $T(z)$ as $T(z) = \sum_n T_n z^n$.

Proposition 1. The number of classes of trees of size n satisfies:

$$T_n = C_n \cdot \sum_{p=1}^{k_n} \left\{ \begin{matrix} n \\ p \end{matrix} \right\} 2^{2n-1-p},$$

where C_n is the number of non labelled binary trees⁴ of size n and $\left\{ \begin{matrix} n \\ p \end{matrix} \right\}$ is the Stirling number of the second kind⁵.

Proof. Once the structure of the binary tree is chosen (factor $2^{n-1}C_n$), we partition the set of leaves into p parts such that two leaves that belong to the same part are labelled by the same variable: it gives the contribution $\left\{ \begin{matrix} n \\ p \end{matrix} \right\}$. Then, we choose to label each leaf by a positive or negative literal (contribution 2^n). The equivalence relationship states that a tree and the one obtained by replacing the positive literals corresponding to a fixed variable by its negative literal (and conversely) are equivalent. Thus, for each class we double-count the number of trees (correction 2^{-p}). \square

Given a set \mathcal{S} of equivalence classes of trees and S_n the number of elements of \mathcal{S} of size n , we define the **ratio** of \mathcal{S} by $\mu_n(\mathcal{S}) = \frac{S_n}{T_n}$. For a given Boolean function f , we denote by $T_n\langle f \rangle$ the number of equivalence classes of trees of size n that compute a function of $\langle f \rangle$, and we define the **probability** of $\langle f \rangle$ as

$$\mathbb{P}_n\langle f \rangle = \frac{T_n\langle f \rangle}{T_n}.$$

One goal of this paper consists in studying the behaviour of the probabilities $(\mathbb{P}_n\langle f \rangle)_{f \in \mathcal{F}}$ when the size n of the trees tends to infinity.

3 Results

We state here our main result: the behaviour of $\mathbb{P}_n\langle f \rangle$ for all fixed function $f \in \mathcal{F}$ in the framework of **and/or** trees. Saying that f is fixed means that its complexity is independent from n .

The main idea of this part is that *a typical tree computing a Boolean function f is a minimal tree of f in which has been plugged a large tree, that does not distort the function computed by the minimal tree.*

⁴ In Proposition 1, C_n is the $(n-1)$ th Catalan number (see e.g. [9, p. 6–7]).

⁵ In Proposition 1, $\left\{ \begin{matrix} n \\ p \end{matrix} \right\}$ is the number of partitions of n objects in p non-empty subsets (see e.g. [9, p. 735–737]).

Definition 3. Let $\langle f \rangle$ be a fixed class of Boolean functions. We denote by $L\langle f \rangle$ (resp. $E\langle f \rangle$) the common complexity (resp. number of essential variables) of the functions of $\langle f \rangle$. The **multiplicity** of the class $\langle f \rangle$, denoted by $R\langle f \rangle$, is the number $L\langle f \rangle - E\langle f \rangle$: it corresponds to the number of repetitions of variables in a minimal tree of $\langle f \rangle$.

Theorem 1. There exists a sequence $(M_n)_{n \geq 1}$ with $M_n \sim \frac{n}{\ln n}$ and such that, for all fixed class $\langle f \rangle$ of Boolean functions, there exists a positive constant $\lambda_{\langle f \rangle}$ such that the probability of $\langle f \rangle$ satisfies, asymptotically when n tends to infinity,

$$\mathbb{P}_n\langle f \rangle = \begin{cases} \sim \lambda_{\langle f \rangle} \cdot \left(\frac{1}{k_{n+1}}\right)^{R\langle f \rangle+1}, & \text{if, for large enough } n, k_n \leq M_n; \\ \sim \lambda_{\langle f \rangle} \cdot \left(\frac{n}{\log n}\right)^{R\langle f \rangle+1} & \text{otherwise.} \end{cases}$$

Let us first remark that the constant $\lambda_{\langle f \rangle}$ is independent from k_n (and from n). This result has been observed, without explanation, in [6,7].

We note also that both constant functions true and false are alone in their respective equivalence classes, and we define their complexity to be 0.

In the *finite* context [2,4], each Boolean function was studied separately instead of being considered among its equivalence class. However, the *finite* context is linked to the particular case of our model where there exists an fixed integer k such that $k_n = k$ for all $n \geq 1$. We can translate the result obtained by Kozik in terms of equivalence classes by summing over all Boolean functions belonging to a given equivalence class: remark that there are $\binom{k}{E\langle f \rangle} 2^{E\langle f \rangle}$ functions in the equivalence class of a given Boolean function f , therefore, the result of Kozik is equivalent to: for all fixed Boolean function $\langle f \rangle$, asymptotically when k tends to infinity,

$$\lim_{n \rightarrow +\infty} \mathbb{P}_{n,k}\langle f \rangle = \Theta\left(\frac{1}{k^{L\langle f \rangle - E\langle f \rangle + 1}}\right) = \Theta\left(\frac{1}{k^{R\langle f \rangle + 1}}\right).$$

Concerning the infinite context [6,7], we notice that the cases such that k_n is larger than n are equivalent to the model $k_n = n$, even if $k_n = \infty$.

4 Technical key points

Next we state the technical core of our results, and we demonstrate how a threshold does appear according to the behaviour of k_n as n tends to infinity.

4.1 Threshold induced by k_n 's behaviour

Definition 4. Let n be a positive integer. The sequence $(a_p)_{p \in \{1, \dots, n\}} = \left(\frac{p^n}{p!} 2^{-p}\right)$ is unimodal. More precisely, there exists a integer M_n such that $(a_p)_p$ is strictly increasing on $\{1, 2, \dots, M_n\}$ and strictly decreasing on $\{M_n + 1, \dots, n\}$.

Note that the sequence (a_p) is related to the terms T_n , cf. Proposition 1. Although the sequence (a_p) is not directly equal to the Stirling numbers of the second kind it is obviously linked to it (cf. next Proposition 2). Therefore we can expect the same kind of behaviour for their maximum (cf. [10,11]).

Lemma 1. *The sequence $(M_n)_n$ is increasing and asymptotically satisfies:*

$$M_n \sim \frac{n}{\ln n}.$$

The proof can be adapted from the approach of Harper [12]. However, simpler arguments are exhibited in Appendix A.

Definition 5. *Let us define the following quantity: $B_{n,k_n} = \sum_{p=1}^{k_n} \binom{n}{p} 2^{-p}$. The number B_{n,k_n} quantitatively represents the labelling constraints of leaves-labelling by variables (cf. Proposition 1).*

Using the following proposition, we will further exhibit bounds on B_{n,k_n} .

Proposition 2 (Comtet, 74). *For all $n \geq 1$, for all $p \in \{1, \dots, n\}$,*

$$\frac{p^n}{p!} - \frac{(p-1)^n}{(p-1)!} \leq \binom{n}{p} \leq \frac{p^n}{p!}.$$

These inequalities can be seen as some specific case of Bonferroni inequalities (see [13, Section 4.7]). For a simpler proof refer to Sibuya [14].

Next lemma is dedicated to understand the asymptotic behaviour of B_{n,k_n} : roughly speaking, before the threshold: $k_n \leq M_n$, B_{n,k_n} is equivalent to its last term, and after M_n , it is equivalent to the sum of a few terms around M_n .

Lemma 2. *Let $(u_n)_n$ be an increasing sequence tending to infinity. Then, asymptotically:*

$$B_{n,u_n} \sim \frac{u_n^n}{u_n!} 2^{-u_n} \quad \text{if } u_n \leq M_n \text{ for large enough } n. \quad (1)$$

$$= \Theta \left(\sum_{p=M_n}^{M_n+\eta_n} \frac{p^n}{p!} 2^{-p} \right) \quad \begin{array}{l} \text{if } u_n \geq M_n \text{ for large enough } n, \\ \text{where } \eta_n = \min\{\ln n, u_n - M_n\}. \end{array} \quad (2)$$

This lemma is proved in Appendix A.

Lemma 3. *Let us assume that $k_n \leq M_n$ for large enough n , then, asymptotically when n tends to infinity,*

$$\frac{B_{n,k_{n+1}}}{B_{n+1,k_{n+1}}} = O\left(\frac{1}{k_{n+1}}\right).$$

Lemma 4. *Let us assume that $k_n \geq M_n$ for large enough n , then, asymptotically when n tends to infinity,*

$$\frac{B_{n,k_{n+1}}}{B_{n+1,k_{n+1}}} = O\left(\frac{\ln n}{n}\right).$$

Definition 6. *Let the ratio rat_n be the quantitative evolution of the leaves-labelling constraints from trees of size n to size $n+1$: $\text{rat}_n = B_{n,k_{n+1}}/B_{n+1,k_{n+1}}$. Its asymptotic behaviour has been quantified in Lemmas 3 and 4.*

4.2 Adjustment of Kozik's pattern language theory

In 2008, Kozik [4] introduced a quite effective way to study Boolean trees: he defined a notion of pattern that permits to easily classify and count large trees according to some constraints on their structure. Kozik applied this pattern theory to study and/or trees with a finite number of variables, but this pattern theory has been extended to different models of Boolean trees (see for example paper [5]).

Let us adapt the definitions of patterns to our new model and then prove extended results of Kozik's paper.

Definition 7. A **pattern language** is a set of binary trees with internal nodes labelled by \wedge or \vee and with external nodes labelled by \bullet or \square . Leaves labelled by \bullet are called **pattern leaves** and leaves labelled by \square are called **placeholders**. Given a pattern language L and a family of trees \mathcal{M} , we denote by $L[\mathcal{M}]$ the family of all trees obtained by replacing every placeholder in an element from L by a tree from \mathcal{M} .

The generating function of a pattern L is $\ell(x, y) = \sum_{d,p} L(d, p) x^d y^p$, where $L(d, p)$ is the number of elements of L with d pattern leaves and p placeholders.

Definition 8. We define the composition of two pattern languages $L[P]$ as the pattern language of trees which are obtained by replacing every placeholder of a tree from L by a tree from P .

Definition 9. A pattern language L is **sub-critical** for a family \mathcal{M} if the generating function $m(z)$ of \mathcal{M} has a square-root singularity τ , and if $\ell(x, y)$ is analytic in some set $\{(x, y) : |x| \leq \tau + \varepsilon, |y| \leq m(\tau) + \varepsilon\}$ for some positive ε .

Definition 10. Given an element of $L[\mathcal{M}]$, its number of **L -repetitions** is the number of its L -pattern leaves minus the number of different variables that appear in the labelling of its L -pattern leaves. The number of its **L -restrictions** is the number of its L -pattern leaves that are labelled by essential variables of the function computed by the tree, plus the number of its L -repetitions.

On the left-hand side of Fig. 1, we have depicted a pattern tree that computes the constant function **true** whatever the placeholder is replaced by. It exhibits one repetition (of the variable x_1) and thus one restriction since the function **true** has no essential variables.

Definition 11. Let \mathcal{I} be the family of the trees with internal nodes labelled by a connective and leaves without labelling, i.e. the family of tree-structures.

The generating function of \mathcal{I} satisfies $I(z) = z + 2I(z)^2$, which implies $I(z) = (1 - \sqrt{1 - 8z})/4$ and its dominant singularity is $1/8$.

The following key-lemma is a generalization of Kozik's one [4, Lemma 3.8]:

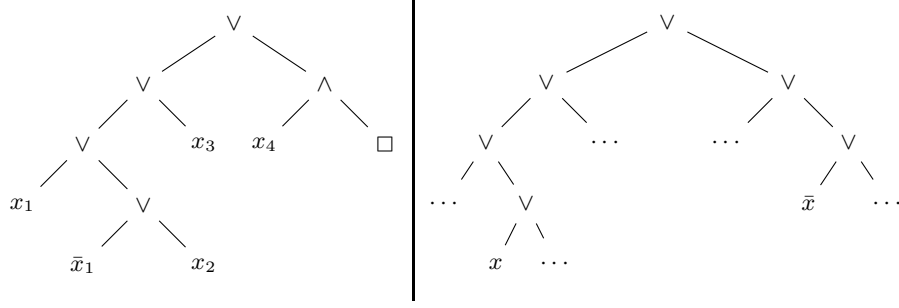


Fig. 1: Left: a pattern tree that computes the function **true**. Right: a simple tautology.

Lemma 5. *Let L be an unambiguous pattern, and \mathcal{T} the families of and/or trees. Let $T_n^{[r]}$ (resp. $T_n^{[\geq r]}$) be the number of labelled (with at most k_n variables) trees of $L[\mathcal{T}]$ of size n and with r L -repetitions (resp. at least r L -repetitions). We assume that L is sub-critical for the family \mathcal{I} of the unlabelled-leaves trees. Then, asymptotically when n tends to infinity,*

$$\frac{T_n^{[r]}}{T_n} = O(\text{rat}_n^r) \quad \text{and} \quad \frac{T_n^{[\geq r]}}{T_n} = O(\text{rat}_n^r).$$

Proof. The number of labelled trees of $L[\mathcal{T}]$ of size n and with at least r L -repetitions is given by:

$$T_n^{[\geq r]} = \sum_{d=r+1}^n I_n(d) \text{Lab}(n, k_n, d, r),$$

where $I_n(d)$ is the number of tree-structures with d L -pattern leaves and the number $\text{Lab}(n, k_n, d, r)$ corresponds to the number of leaves-labellings of these trees giving at least r L -repetitions. The following enumeration contains some double-counting and we therefore get an upper bound:

$$\text{Lab}(n, k_n, d, r) \leq 2^n \cdot \sum_{j=1}^r \binom{d}{r+j} \left\{ \begin{matrix} r+j \\ j \end{matrix} \right\} B_{n-r-j+1, k_n}.$$

The factor 2^n corresponds to the polarity of each leaf (the literal labelling is positive or negative); the index j stands for the number of different variables involved in the r repetitions; the binomial factor chooses the pattern leaves that are involved in the r repetitions; the Stirling number partition splits $r+j$ leaves into j parts; finally, the factor $B_{n-r-j+1, k_n}$ chooses which variable is assigned to each class of leaves. Therefore,

$$T_n^{[\geq r]} \leq 2^n \cdot B_{n-r, k_n} \sum_{j=1}^r \left\{ \begin{matrix} r+j \\ j \end{matrix} \right\} \sum_{d=r+j}^n I_n(d) \binom{d}{r+j}.$$

Let $\ell(x, y)$ be the generating function of the pattern L . Then, for all $p \geq 0$,

$$\frac{z^p}{p!} \frac{\partial^p \ell}{\partial x^p}(z, I(z)) = \sum_{n=1}^{\infty} \sum_{d=1}^{\infty} I_n(d) \binom{d}{p} z^n.$$

Thus,

$$\frac{T_n^{[\geq r]}}{T_{n, k_n}} \leq \frac{B_{n-r, k_n}}{B_{n, k_n}} \sum_{j=1}^r \begin{Bmatrix} r+j \\ j \end{Bmatrix} \frac{[z^n] z^{r+j} \frac{\partial^{r+j} \ell}{\partial x^{r+j}}(z, I(z))}{[z^n] I(z)}.$$

Since $z^{r+j} \frac{\partial^{r+j} \ell}{\partial x^{r+j}}(z, I(z))$ and $I(z)$ have the same singularity because of the sub-criticality of the pattern L according to \mathcal{I} , the previous sum is constant when n tends to infinity and so we conclude:

$$\frac{T_n^{[r]}}{T_n} \leq \frac{T_n^{[\geq r]}}{T_n} = \Theta\left(\frac{B_{n-r, k_n}}{B_{n, k_n}}\right) = O(\text{rat}_n^r).$$

□

5 Behaviour of the probability distribution

Now that we have adapted the pattern theory to our model, we are ready to quantitatively study it. A first step is to understand the asymptotic behaviour of $\mathbb{P}_n\langle \text{true} \rangle$. It is indeed natural to focus on this “simple” function before considering a general class $\langle f \rangle$; and moreover, it happens to be essential for the continuation of the study. In addition, the methods used to study tautologies (mainly pattern theory) will also be the core of the proof for a general equivalence class. We prove in this section the main Theorem 1 for both classes $\langle \text{true} \rangle$ and $\langle \text{false} \rangle$ of complexity zero, using the duality of both connectives \wedge and \vee and both positive and negative literals. The main ideas of the proof for a general equivalence class will be detailed in Section 5.2, but the details will be postponed into Appendix C.

5.1 Tautologies

Let us recall that a **tautology** is a tree that represents the Boolean function **true**. Let us consider the family \mathcal{A} of tautologies. In this part, we prove that the probability of $\langle \text{true} \rangle$ is equivalent to the ratio of a simple subset of tautologies.

Definition 12 (cf. right-hand side of Fig. 1). *A simple tautology is an and/or tree that contains two leaves labelled by a variable x and its negation \bar{x} and such that all internal nodes from the root to both leaves are labelled by \vee -connectives. We denote by ST the family of simple tautologies.*

In order to prove Theorem 1 for the class $\langle \text{true} \rangle$ and even to give the more precise result $\mathbb{P}_n\langle \text{true} \rangle \sim 3/4 \cdot \text{rat}_n$, first we compute the ratio of simple tautologies.

Lemma 6. *The ratio of simple tautologies verifies*

$$\mu_n(ST) = \frac{ST_n}{T_n} \sim \frac{3}{4} \text{rat}_n, \text{ when } n \text{ tends to infinity.}$$

Moreover, asymptotically when n tends to infinity, almost all tautologies are simple tautologies.

Proof. The proof is divided in two steps. The first one is dedicated to the computation of the ratio $\mu_n(ST)$. The second part of the proof shows that almost all tautologies are simple tautologies.

Let us consider the non-ambiguous pattern language $S = \bullet |S \vee S| \square \wedge \square$. Remark that a tree, such that two S -pattern leaves are labelled by a variable and its negation, is a simple tautology. The generating function of S is $s(x, y) = \frac{1}{2}(1 - \sqrt{1 - 4(x + y^2)})$. It is sub-critical for \mathcal{I} . The generating function $\tilde{I}(z) = \frac{1}{2} \partial^2 / \partial x^2 (s(xz, I(z)))|_{x=1}$ enumerates and/or trees with two marked distinct leaves. Therefore, $2^{n-1} \tilde{I}_n B_{n-1}$ is the number of simple tautologies where we count twice simple tautologies realized simultaneously by two pairs of leaves. The ratio of this family, with double-counting and denoted by DC , is given by

$$\mu_n(DC) = \frac{2^{n-1} \tilde{I}_n B_{n-1, k_n}}{2^{n+1} I_n B_{n, k_n}},$$

and using a consequence of [9, Theorem VII.8] (cf. a detailed proof in [7]):

$$\lim_{n \rightarrow \infty} \frac{\tilde{I}_n}{I_n} = \lim_{z \rightarrow \frac{1}{8}} \frac{\tilde{I}'(z)}{I'(z)} = 3.$$

Thus, we get the upper bound $\frac{3}{4} \text{rat}_n$ for the ratio of simple tautologies.

It remains to deal with the double-counting in order to compute a lower bound. In the family DC , simple tautologies, realized by a unique pair of leaves, are counted once, those that are realized by two pairs of leaves are counted twice, and so on. Let us denote by ST^i the family of simple tautologies counted exactly i times. Inclusion-exclusion principle gives: $ST_n = DC - \sum_{i \geq 1} (-1)^i \cdot ST_n^i$. Moreover, it can be seen that a tree in ST^2 (resp. in ST^3 , resp. ST^i) has at least 2 (resp. 3, resp. $\lfloor 2\sqrt{i} - 1 \rfloor$) S -repetitions. Therefore, by Lemma 5, the ratio of the family ST^i is:

$$\mu_n(ST^i) = O\left((\text{rat}_n)^{2\sqrt{i}-1}\right), \text{ when } n \text{ tends to infinity.}$$

$$\begin{aligned} \text{Thus, } \mu_n(DC) - \mu_n(ST) &= \left| \sum_{i=2}^n (-1)^{i+1} \mu_n(ST^i) \right| \leq \sum_{i=2}^3 \mu_n(ST^i) + \sum_{i=4}^n \mu_n(ST^i) \\ &\leq O\left(\left(\frac{\ln n}{n}\right)^2\right) + n \cdot O\left(\left(\frac{\ln n}{n}\right)^3\right) = O\left(\left(\frac{\ln n}{n}\right)^2\right). \end{aligned}$$

Consequently, asymptotically, $\mu_n(ST) = \mu_n(DC) + o(\text{rat}_n) \sim \frac{3}{4} \cdot \text{rat}_n$.

Let us now turn to the second part of the proof: asymptotically, almost all tautologies are simple tautologies. Let us consider the pattern $N = \bullet | N \vee N | \square \wedge N$. This pattern is unambiguous, its generating function verifies $n(x, y) = x + n(x, y)^2 + y \cdot n(x, y)$ and is thus equal to $\frac{1}{2}(1 - y - \sqrt{(1 - y)^2 - 4x})$. It implies that N is sub-critical for the family \mathcal{I} of tree-structures.

A tautology has at least one $N[N]$ -repetition, otherwise, we can assign all its N -pattern leaves to false and, the whole tree computes false: impossible for a tautology.

Consider a tautology t with exactly one $N[N]$ -repetition. this repetition must be a $x|\bar{x}$ repetition and must occur among the N -pattern leaves, using the same kind of argument than above.

Then, let us assume that there is an \wedge -node denoted by ν between the N -pattern leaf x and the root of the tree. This node ν has a left subtree t_1 and a right subtree t_2 . Assume that the leaf x appears in t_1 . Then, one can assign all the N -pattern leaves of t_2 (which are $N[N]$ -pattern leaves of t) to false, since there is no more repetition among the $N[N]$ -pattern leaves of t . Also assign all the pattern leaves of t minus the subtree rooted at ν to false. Then, we can see that t computes false: impossible. We have thus shown that t is a simple tautology.

Finally, tautologies with exactly one $N[N]$ -repetition are simple tautologies, a tautology must have at least one $N[N]$ -repetition and, thanks to Lemma 5, tautologies with more than one $N[N]$ -repetitions have a ratio of order $o(\text{rat}_n)$, which is negligible in front of the ratio of simple tautologies. \square

5.2 Probability of a general class of functions

With similar arguments than those used for tautologies, we prove that the probability of the class of projections (i.e. $(x_i)_{i \geq 1} \mapsto x_j$) is equivalent to $5/8 \cdot \text{rat}_n$. The proof is detailed in Appendix B.

Let us turn now to the general result: the behaviour of $\mathbb{P}_n\langle f \rangle$ for all fixed $f \in \mathcal{F}$. The main idea of this part is that, roughly speaking, *a typical tree computing a Boolean function in $\langle f \rangle$ is a minimal tree of $\langle f \rangle$ in which has been plugged a single large tree*. The goal of this section is to give the main ideas of the proof of Theorem 1, the complete proof is given in Appendix C.

Proof (sketch). For a given class of Boolean functions $\langle f \rangle$ our goal is to obtain an asymptotic equivalent to $\mathbb{P}_n\langle f \rangle$.

- We first define several notions of *expansions* of a tree: the idea is to replace in a tree, a subtree S by $T \wedge S$, where T is chosen such that the expanded tree still computes the same function.
- The ratio of minimal trees of $\langle f \rangle$ expanded once is of the order of $\text{rat}_n^{R(f)+1}$.
- The ratio of trees computing a function from $\langle f \rangle$ is equivalent to the ratio of minimal trees expanded once.

The most technical part of the proof is the last one, because we need a precise upper bound of $\mathbb{P}_n\langle f \rangle$. But the ideas are more or less the same as those developed for the class $\langle \text{true} \rangle$. \square

6 Conclusion

We studied a new model of **and/or** trees which is the first one (to our knowledge) to allow the number of variables to depend on the size of the trees into consideration.

Choosing the context of **and/or** trees let us to generalize the powerful Kozik's pattern theory, but we are convinced that all our results also hold in implicational models or in non-binary or non-plane models. Indeed, the key idea is that *each repetition induces a factor rat_n* , and this remains true in all those models – although pattern theory does not adapt to every model, e.g. models with *implication*. Extending our results to these models would give nice unifications of the known results of the literature: papers [4,8,7] and [15,5].

References

1. Lefmann, H., Savický, P.: Some typical properties of large And/Or Boolean formulas. *Random Structures and Algorithms* **10** (1997) 337–351
2. Chauvin, B., Flajolet, P., Gardy, D., Gittenberger, B.: And/Or trees revisited. *Combinatorics, Probability and Computing* **13**(4–5) (2004) 475–497
3. Gardy, D.: Random Boolean expressions. In: *Colloquium on Computational Logic and Applications*. Volume AF., DMTCS (2006) 1–36
4. Kozik, J.: Subcritical pattern languages for And/Or trees. In: *Fifth Colloquium on Mathematics and Computer Science*, DMTCS Proceedings (2008)
5. Genitrini, A., Gittenberger, B., Kraus, V., Mailler, C.: Associative and commutative tree representations for boolean functions. Submitted to *Journal: Combinatorics Probability & Computing*
6. Genitrini, A., Kozik, J., Zaionc, M.: Intuitionistic vs. classical tautologies, quantitative comparison. In: *TYPES*. (2007) 100–109
7. Genitrini, A., Kozik, J.: In the full propositional logic, 5/8 of classical tautologies are intuitionistically valid. *Ann. of Pure and Applied Logic* **163**(7) (2012) 875–887
8. Fournier, H., Gardy, D., Genitrini, A., Gittenberger, B.: The fraction of large random trees representing a given boolean function in implicational logic. *Random Structures and Algorithms* **40**(3) (2012) 317–349
9. Flajolet, P., Sedgewick, R.: *Analytic Combinatorics*. Cambridge U.P. (2009)
10. Dobson, A.J.: A note on stirling number of the second kind. *Combinatorial Society* **5** (1968) 212–214
11. Canfield, E.R., Pomerance, C.: On the problem of uniqueness for the maximum stirling number(s) of the second kind. *INTEGERS (electronic journal)* **2** (2002)
12. Harper, L.H.: Stirling behavior is asymptotically normal. *Ann. Math. Stat.* **38** (1966) 410–414
13. Comtet, L.: *Advanced Combinatorics: The Art of Finite and Infinite Expansions*. Reidel (1974)
14. Sibuya, M.: Log-concavity of Stirling numbers and unimodality of Stirling distributions. *Ann. of the Institute of Statistical Mathematics* **40**(4) (1988) 693–714
15. Genitrini, A., Gittenberger, B., Kraus, V., Mailler, C.: Probabilities of Boolean functions given by random implicational formulas. *Electronic Journal of Combinatorics* **19**(2) (2012) P37, 20 pages (electronic)

A Proofs of the technical core

Proof (of Lemma 1). Let p be an integer in $\{1, \dots, n-1\}$. By Definition 4,

$$\frac{a_{p+1}}{a_p} = \left(\frac{p+1}{p}\right)^n \frac{1}{2(p+1)},$$

and consequently, for large enough n ,

$$\frac{a_{p+1}}{a_p} > 1 \iff n \ln \left(\frac{p+1}{p}\right) - \ln(2(p+1)) > 0.$$

The function $(p \mapsto n \ln \left(\frac{p+1}{p}\right) - \ln(2(p+1)))$ is strictly decreasing. Since it tends to $+\infty$ at $p = 1$ and to $-\infty$ at $p = n-1$, both when n tends to infinity, there exists a unique M_n such that (a_p) is strictly increasing on $\{1, \dots, M_n\}$ and strictly decreasing on $\{M_n+1, \dots, n\}$.

Let us denote by x_n the single solution of equation:

$$\left(\frac{x+1}{x}\right)^n \frac{1}{2(x+1)} = 1. \quad (3)$$

Since, asymptotically when n tends to infinity,

$$\left(\frac{\frac{n}{\ln n} + 1}{\frac{n}{\ln n}}\right)^n \frac{1}{2(\frac{n}{\ln n} + 1)} \sim \frac{\ln n}{2},$$

we have that $n/\ln n \leq x_n$ and therefore, x_n tends to infinity. Thus, Equation (3) evaluated in x_n is equivalent to

$$n \ln \left(1 + \frac{1}{x_n}\right) = \ln 2 + \ln(x_n + 1),$$

which implies $x_n \ln x_n \sim n$ when n tends to infinity. We easily deduce from this asymptotic relation that $\ln x_n \sim \ln n$ and that $x_n \sim \frac{n}{\ln n}$ when n tends to infinity. Since $M_n = \lfloor x_n \rfloor$, we conclude that $M_n \sim n/\ln n$ when n tends to infinity. \square

In view of Proposition 2, we have the following bounds:

$$\frac{1}{2} \cdot \sum_{p=1}^{u_n-1} \frac{p^n}{p! 2^p} + \frac{u_n^n}{u_n! 2^{u_n}} \leq B_{n,u_n} \leq \sum_{p=1}^{u_n} \frac{p^n}{p! 2^p}. \quad (4)$$

Proof (of Lemma 2, assertion (1)). When $u_n \leq M_n$, for large enough n , let us prove that both bounds of Equation (4) are of the same order as n tends to infinity. Let us first prove that both bounds are equivalent to their last (and common) term, namely $u_n^n \cdot (u_n! 2^{u_n})^{-1}$.

Let us denote by S_{u_n-1} , the sum $\sum_{p=1}^{u_n-1} a_p$. Let ε be positive. We define δ_n as the minimum value between $u_n - 1$ and $(\ln n)^{1-\varepsilon}$. We divide the sum S_{u_n-1}

in two parts: the last δ_n terms and the other ones (if they exist). Let us recall that $(a_p)_{p \geq 1}$ is increasing while $p \leq M_n$. It implies

$$\frac{S_{u_n-1}}{a_{u_n}} \leq \delta_n \cdot \frac{a_{u_n-1}}{a_{u_n}} + u_n \cdot \frac{a_{u_n-1-\delta_n}}{a_{u_n}}.$$

Let us first focus on the following factor:

$$\frac{a_{u_n-1}}{a_{u_n}} = 2u_n \cdot \left(1 - \frac{1}{u_n}\right)^n \leq \frac{2}{\ln n}.$$

Since $\delta_n \leq (\ln n)^{1-\varepsilon}$, we have that $\delta_n \cdot \frac{a_{u_n-1}}{a_{u_n}} \leq \frac{2}{(\ln n)^\varepsilon}$.

If $\delta_n = u_n - 1$, then S_{u_n-1} is negligible in front of a_{u_n} . Otherwise, if $\delta_n = (\ln n)^{1-\varepsilon}$, then

$$\frac{a_{u_n-1-\delta}}{a_{u_n}} \lesssim (2u_n)^{1+\sqrt{\ln n}} \cdot \left(1 - \frac{1+\sqrt{\ln n}}{u_n}\right)^n \lesssim \left(\frac{2}{\ln n}\right)^{1+\sqrt{\ln n}},$$

where $a_n \lesssim b_n$ means that a_n is smaller than a quantity equivalent to b_n , asymptotically when n tends to infinity. Thus

$$\frac{S_{u_n-1}}{a_{u_n}} \leq n \cdot \left(\frac{2}{\ln n}\right)^{1+\sqrt{\ln n}} + \sqrt{\ln n} \cdot \frac{2}{\ln n} \xrightarrow{n \rightarrow +\infty} 0.$$

So S_{u_n-1} is negligible in front of a_{u_n} in that case too. Finally, B_{n,u_n} is equivalent to a_{u_n} , when n tends to infinity. \square

Proof (of Lemma 2, assertion (2)). In the case $u_n \geq M_n$, it seems not true that both sums are equivalent to the larger term a_{M_n} . However, such a precise result is not necessary..

Let $\eta_n = \min\{u_n - M_n, \ln n\}$. In both bounds of Equation (4), we separate the sums in three parts: the first one from indices 1 to $M_n - 1$; the second one from M_n to $M_n + \eta_n$; and the third one from $M_n + \eta_n + 1$ to u_n (this last sum can eventually be empty).

Using assertion (1) of Lemma 2 and the fact that the second part of the sum contains the term a_{M_n} , we conclude that the first part of the sum is negligible. Let us now prove that the third part (when it is not empty: i.e. $\eta_n = \ln n$) is negligible too, in front of the second.

Let us denote t_n the third part of the sum divided by a_{M_n} . Since (a_p) is decreasing when $p \geq M_n + 1$, we get:

$$t_n = \sum_{p=M_n+\eta_n+1}^{u_n} \frac{p^n}{p!} 2^{-p} \frac{M_n!}{M_n^n} 2^{M_n} \leq u_n \frac{(M_n + \eta_n)^n}{(M_n + \eta_n)!} 2^{-M_n - \eta_n} \frac{M_n!}{M_n^n} 2^{M_n}.$$

Thus, the Stirling formula gives

$$t_n \leq \exp \left(\ln u_n + \eta_n (1 - \ln(2)) + (\eta_n - M_n) \ln \left(1 + \frac{\eta_n}{M_n} \right) - \eta_n \ln(M_n + \eta_n) \right).$$

Since $\eta_n = \Theta(\ln n)$ and $M_n \sim n/\ln n$, we get:

$$t_n \lesssim \exp(-\eta_n \ln n + \eta_n \ln(\ln n)) \xrightarrow{n \rightarrow +\infty} 0.$$

Thus the third part of the sum is negligible and, as n tends to infinity,

$$B_{n,u_n} = \Theta \left(\sum_{p=M_n}^{M_n+\eta_n} \frac{p^n}{p! 2^p} \right).$$

□

Proof (of Lemma 3). In view of Lemma 2 applied to $u_n = k_{n+1}$, if $k_{n+1} \leq M_n$, we get

$$\frac{B_{n,k_{n+1}}}{B_{n+1,k_{n+1}}} \sim \frac{k_{n+1}^n}{k_{n+1}^{n+1}} \frac{k_{n+1}!}{k_{n+1}!} 2^{k_{n+1}-k_{n+1}} \sim \frac{1}{k_{n+1}}.$$

Otherwise, using Lemma 2, when $M_{n+1} \geq k_{n+1} > M_n$, there exists a constant α such that

$$\begin{aligned} \frac{B_{n,k_{n+1}}}{B_{n+1,k_{n+1}}} &\lesssim \alpha \frac{\ln n}{k_{n+1}} \left(\frac{M_n}{k_{n+1}} \right)^{n-M_n-1/2} k_{n+1}^{k_{n+1}-M_n} \\ &\lesssim \alpha \frac{\ln n}{k_{n+1}} \left(\frac{M_n}{M_n + \eta_n} \right)^{n-M_n-1/2} M_{n+1}^{M_{n+1}-M_n}. \end{aligned}$$

since $M_n + \eta_n \leq k_{n+1}$ by definition of η_n , and $k_{n+1} \leq M_{n+1}$ by assumption. Therefore,

$$\frac{B_{n,k_{n+1}}}{B_{n+1,k_{n+1}}} \lesssim \alpha \frac{1}{k_{n+1}}.$$

and Lemma 3 is proved. □

Proof (of Lemma 4). We have $k_{n+1} \geq M_{n+1}$, so $k_{n+1} \geq M_n$. We thus apply Lemma 2, assertion (2) with $u_n = k_{n+1}$. Consequently, there exists a constant α , for large enough n , such that:

$$\begin{aligned} \frac{B_{n,k_{n+1}}}{B_{n+1,k_{n+1}}} &\leq \alpha 2^{M_{n+1}-M_n+\eta_{n+1}} \frac{\eta_n}{\eta_{n+1}} \frac{M_n^n}{(M_{n+1} + \eta_{n+1})^{n+1}} \frac{(M_{n+1} + \eta_{n+1})!}{M_n!} \\ &\lesssim \alpha 2^{M_{n+1}-M_n} \frac{\eta_n}{\eta_{n+1}} \frac{M_n^n}{M_{n+1}^{n+1}} \frac{M_{n+1}!}{M_n!}. \end{aligned}$$

Using the Stirling formula, both properties of Lemma 1, and the fact that η_n/η_{n+1} tends to 1, we conclude that

$$\frac{B_{n,k_{n+1}}}{B_{n+1,k_{n+1}}} = \mathcal{O} \left(\frac{\ln n}{n} \right),$$

and the stated result is proved. □

B Probability of the class of projections

Studying the probability of **true** is essential to understand the model while studying the projections is not necessary. However, it permits to be more familiar with the model and often permits to conjecture the general behaviour of $\mathbb{P}_n\langle f \rangle$. This gives a sufficient reason to deeply study $\mathbb{P}_n\langle x \rangle$ (x is a literal). We will not detail all the proofs that are very similar to those of Section 5.

To calculate the probability of the class of projections we will follow the ideas presented for tautologies: we define a set of trees of simple shape that compute the projection x and call such trees “simple- x ” and then show that the ratio of simple- x is, asymptotically when the size of the trees n tends to infinity, equal to the probability of the projection.

Definition 13 (cf. Figure 2). A **simple- x of type T** is a tree with one subtree reduced to a single leaf and the other subtree being a simple tautology if the root’s label is \wedge or a simple contradiction if the root’s label is \vee .

A **simple- x of type X** is a tree with one subtree reduced to a single leaf ℓ , the root labelled by \wedge (resp. \vee) and the other subtree such that there exists a leaf labelled by the same literal as ℓ linked to the root by a \vee -only path.

We denote by \mathcal{X} the family of simple- x .

Obviously, simple- x are computing the projection x .

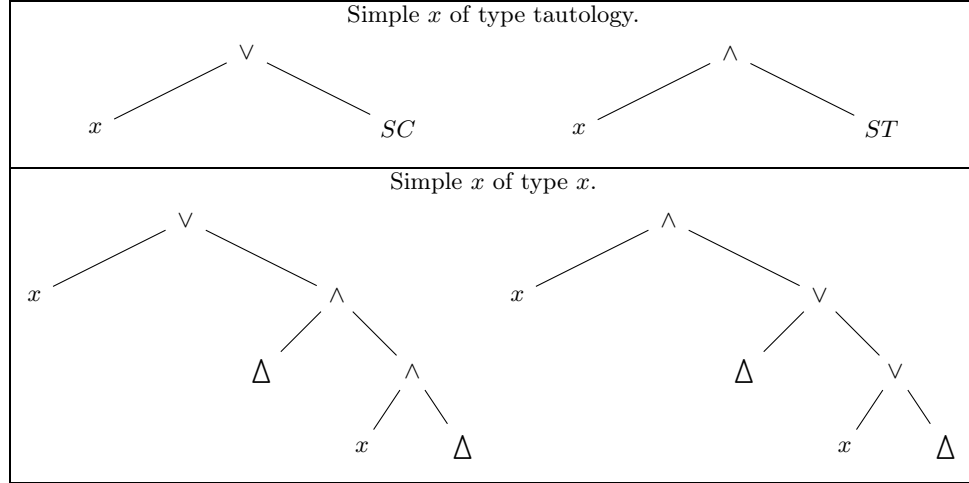


Fig. 2: Examples of simple- x .

Lemma 7. If X_n^T is the number of type T simple- x of size n , we have, when n tends to infinity:

$$\lim_{n \rightarrow +\infty} \frac{X_n^T}{T_n} \sim \frac{3}{8} \text{rat}_n.$$

Proof. We have:

$$\frac{X_n^T}{T_n} \sim \frac{4 \cdot 2^{n-1} B_{n-1, k_n} [z^{n-1}] \frac{\partial^2}{2\partial x^2} s(zx, I(z))|_{x=1}}{T_n}$$

because a type T simple- x of size n is either a tree rooted by \wedge or a tree rooted by \vee (which gives a factor 2), with either its right or its left subtree being a single leaf (which also gives a factor 2), and the other subtree being a simple tautology or a simple contradiction (depending on the root's label) of size $n-1$. Remark that this equation is only true asymptotically when n tends to infinity, since we do double-counting which becomes negligible when n tends to infinity. Thus, asymptotically when n tends to infinity,

$$\frac{X_n^T}{T_{n, k_n}} \sim \frac{4 \cdot 2^{n-1} B_{n-1, k_n} [z^{n-1}] \frac{\partial^2}{2\partial x^2} s(zx, I(z))|_{x=1}}{2^n B_{n, k_n} I_n} = \frac{2 \cdot 2^{n-1} B_{n-1, k_n} \tilde{I}_{n-1}}{2^n B_{n, k_n} I_n}.$$

We already have proved: $\tilde{I}_n/I_n \sim 3$, and $I_{n-1}/I_n = 1/8$, so the result is proved. \square

Lemma 8. *If X_n^X is the number of type X simple- x of size n , we have, asymptotically when n tends to infinity,*

$$\lim_{n \rightarrow +\infty} \frac{X_n^X}{T_n} \sim \frac{\text{rat}_n}{4}.$$

Proof. We have:

$$\frac{X_n^X}{T_n} \sim \frac{4 \cdot 2^{n-1} B_{n-1, k_n} [z^{n-1}] \frac{\partial}{\partial x} s(zx, I(z))|_{x=1}}{2^n B_{n, k_n} I_n}$$

because a type T simple- x of size n is either a tree rooted by \wedge or a tree rooted by \vee (which gives a factor 2), with either its right or its left subtree being a single leaf (which also gives a factor 2), and because the other subtree is a tree where we have chosen one S pattern leaf and labelled it by the same labelled as the first level leaf. Since there can be several S pattern leaves that can have simultaneously the same label as the leaf subtree, we do double counting, but once again, thanks to Lemma 5, this double counting becomes negligible when n tends to infinity. Thus,

$$\frac{X_n^X}{T_n} \sim \frac{4 \cdot 2^{n-1} B_{n-1, k_n}}{2^n B_{n, k_n}} \frac{1}{8}.$$

Since $[z^{n-1}] \frac{\partial}{\partial x} s(zx, I(z))|_{x=1}/I_n \sim 1$ and $I_{n-1}/I_n \sim 1/8$, we get the result. \square

Lemma 9. *Asymptotically when n tends to infinity, the ratio of simple- x is equal to the probability of the projection.*

The proof of this lemma is very similar to the proof of Lemma 6.

C Probability of a general class of Boolean functions

In the following, $\langle f \rangle$ is fixed (and f is one of its representative). T is an and/or tree computing f . Moreover, we will need to consider the patterns $R = N^{(r+1)}[N \oplus P]$ and $\bar{R} = N^{(r+1)}[(N \oplus P)^2]$. Note that the language $N \oplus P$ is defined such that the $N \oplus P$ -pattern leaves of a tree are its N -pattern leaves plus its P pattern leaves. It is proved in [4] that this pattern language is indeed non-ambiguous and sub-critical for I if N and P are.

Proposition 3. *A tree t computing f with at least one leaf on the $(r+2)^{th}$ level of the R pattern must have at least $R(f) + 1$ R -repetitions.*

Proof. Let us assume that t computes f , has at least one leaf on the $(r+2)^{th}$ level of the R pattern but have less than $R(f)$ R -repetitions. Let i be the smallest integer (smaller than $r+2$) such that the number of $N^{(i)}$ -restrictions is equal to the number of $N^{(i-1)}$ -restrictions.

There must be either a repetition or an essential variable in the first level: if there is none, then we can assign all the N pattern leaves to **false** and this operation does not changes the calculated function. The calculated function is then the constant function **false**, which is impossible; so $i \leq r+1$.

First Case: Let us assume that there are strictly less than r $N^{(i)}$ -restrictions. There is no repetition and no essential variable in the pattern leaves at level i . Therefore, we can assign them all to **false** and make the placeholders of the level $i-1$ compute **false**. Let us replace those placeholders by **false** in the tree. Furthermore, replace by **false** all the non-essential remaining variables. And simplify the obtained tree to simplify all the constant leaves **false** and **true**. We obtain a tree t^* , which still computes f , and whose leaves are all former $N^{(i-1)}$ pattern leaves of t labelled by essential variables. The tree t^* therefore contains strictly less than r leaves, which is impossible since the complexity of f is r .

Second Case: Let us assume that t has exactly r $N^{(i)}$ -restrictions. Since $i \leq r+1$, there is no restriction in the placeholders of the level $r+2$. Therefore, we can replace the placeholders by wildcards \star , which means that those wildcards can be evaluated to **true** or **false** independently from each other and without changing the function computed by t . We can also replace the remaining leaves labelled by non-essential and non-repeated variables by such wildcards.

We simplify those wildcards. Such a simplification has to delete at least one non-wildcard leaf. If we deleted a non-repeated essential variable, then the tree t^* does not depend on this essential variable and computes f : this is impossible. Thus, we deleted a repetition: t^* has strictly less than $R(f)$ repetitions and computes f . It is impossible. \square

Remark that in Lemma 5, we only count repetitions and not restrictions as it was done in the original Lemma by Kozik. Because in terms of equivalence classes, essential variables are no longer relevant. Though, we will need to consider essential variables and the following lemma permits to handle them.

Lemma 10. *Let L be an unambiguous pattern, sub-critical for \mathcal{I} . Let f be a fixed Boolean functions and \mathcal{M}_f the set of minimal trees computing f . Let \mathcal{E} be the family of trees obtained by expanding once a tree of \mathcal{M}_f by trees having exactly p L -restrictions. Then,*

$$\mu_n(\mathcal{E}) \sim \alpha \cdot \text{rat}_n^{R(f)+p},$$

with $\alpha > 0$ a constant.

Proof. Let E_n be the number of trees of size n in \mathcal{E} . We will denote by i the number of leaves that are involved in the p L -restrictions of the expansion tree: i is at least $p+1$ and at most $2p$. With negligible double-counting,

$$\mu_n(\mathcal{E}) = \frac{E_n}{T_n} = \sum_{i=p+1}^{2p} [z^{n-L(f)}] \frac{\partial^i}{i! \partial x^i} (\ell(xz, I(z)))|_{x=1} \frac{2^n B_{n-p-R(f), k_n}}{2^n I_n B_{n, k_n}}.$$

Since L is sub-critical for \mathcal{A} ,

$$\sum_{i=p+1}^{2p} \frac{[z^{n-L(f)}] \partial^i / i! \partial x^i (\ell(xz, I(z)))|_{x=1}}{I_n} \sim \alpha \cdot \frac{I_{n-L(f)}}{I_n} \sim \left(\frac{1}{8}\right)^{L(f)} > 0$$

asymptotically when n tends to infinity. Therefore, in view of Section 4,

$$\mu_n(\mathcal{E}) \sim \alpha \cdot \text{rat}_n^{R(f)+p}.$$

□

Consider the family of trees obtained by replacing a subtree s by $s \wedge t_e$ where t_e is a simple tautology into a minimal tree of f . Let us denote by E_n the number of such trees of size n . Since a simple tautology has at least one S -restriction, thanks to 10,

$$\frac{E_n}{T_n} \sim \alpha \cdot \text{rat}_n^{R(f)+1}.$$

Thanks to Lemma 5, we know that terms computing f with more than $R(f)+2$ repetitions are negligible in front of the above family. Therefore, since trees with no leaf on the $(r+2)^{\text{th}}$ level are negligible, we proved Theorem 1.

In fact, we can show a more precise result:

Theorem 2. *Let f be a fixed Boolean function, then, asymptotically when n tends to infinity,*

$$\mathbb{P}_n\langle f \rangle \sim \lambda_{\langle f \rangle} \text{rat}_n^{R(f)+1},$$

where $\lambda_{\langle f \rangle}$ is a positive constant.

The key point of the proof of this Theorem is that a typical tree computing a function from $\langle f \rangle$ is a minimal tree of this function which has been expanded once. In the following, we will only consider two different expansions:

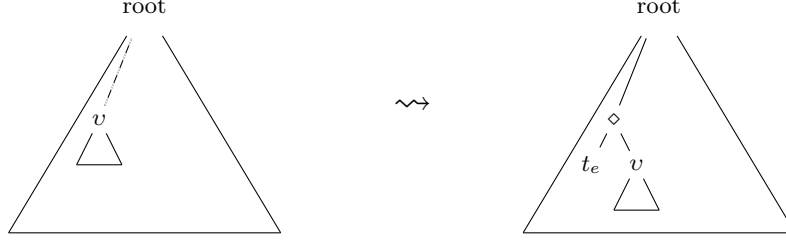


Fig. 3: An expansion at node v . Note that the expansion tree t_e could have been on the right size of the \diamond -connective instead of its left side.

Definition 14 (cf. Figure 3). Recall that an **expansion** of a tree t is a tree obtained by replacing a subtree s of t by $s \diamond t_e$ (or $t_e \diamond s$) where $\diamond \in \{\wedge, \vee\}$.

An expansion is a **T-expansion** if the expansion tree t_e is a simple tautology and the connective \diamond is \wedge (or a simple contradiction and the connective \diamond is \vee).

An expansion is a **X-expansion** if the expansion tree t_e has a leaf linked to the root by a \wedge -path (resp. a \vee -path) and the \diamond connective is a \vee (resp. \wedge).

Lemma 11. The ratio of minimal trees of f expanded once verifies, asymptotically when n tends to infinity

$$\mu_n(E[\mathcal{M}_f]) = \alpha \cdot \text{rat}_n^{R(f)+1} + o\left(\text{rat}_n^{R(f)+1}\right).$$

This lemma is a direct consequence of Lemma 10.

Lemma 12. Let f be a fixed Boolean function and let \mathcal{M}_f be the set of minimal trees of f .

$$\mathbb{P}_n\langle f \rangle \sim \mu_n(E[\mathcal{M}_f]) \text{ when } n \rightarrow +\infty.$$

Proof. Let t be a term computing f . Such a term must have at least $R(f) + 1$ \bar{R} -repetitions. Moreover, thanks to Lemma 5, trees with at least $R(f) + 2$ \bar{R} -repetitions are negligible. We will show that a tree with exactly $R(f) + 1$ \bar{R} -repetitions is in fact a minimal tree expanded once.

The term t must also have $R(f) + 1$ R -repetitions and therefore, there is no additional repetition when we consider the $(r + 3)^{\text{st}}$ level of the \bar{R} -pattern.

Let i be the first level such that the number of $N^{(i)}$ restrictions is equal to the number of $N^{(i-1)}$ -restrictions. Since there must be a restriction on the first level, $i \leq r + 1$.

First Case: Assume that an essential variable α appears on the pattern leaves of the $(r + 3)^{\text{th}}$ level. Therefore, t has at most $L(f)$ $N^{(i)}$ -restrictions. Let us replace the placeholders of the $(i - 1)^{\text{th}}$ level by **false** and assign all the remaining non-essential variables to **false**. Simplify the tree to obtain a new and/or tree denoted

by t^* . The leaves of this tree are former $N^{(i-1)}$ -pattern leaves of t , labelled by essential variables and t^* still computes f . But the variable α is essential for f : thus it must still appear in the leaves of t^* , and by deleting its occurrence in the leaves of the $(r+3)^{\text{th}}$ level, we deleted one repetition. Therefore, t^* has at most $L(f) - 1$ leaves which is impossible!

Second Case: There is no essential variable among the the pattern leaves of the $(r+3)^{\text{th}}$ level. Since there is also no repetition at this level, we can replace the placeholders of the level $(r+3)$ to wildcards. We also replace the remaining non essential and non-repeated variables by wildcards. We then simplify the wildcards and obtained a simplified tree t^* , computing f , with no wildcards and which leaves are former leaves of the trees t , essential or repeated. During the simplification process, we have deleted at least one of these leaves and therefore t^* has at most $L(f)$ leaves: it is a minimal tree of f .

Let us consider the following fact: The lowest common ancestor of all the wildcards in t has been suppressed during the simplification process.

Assume that this fact is false: then two wildcards have been simplified independently during the simplification process, and thus, at least two essential or repeated variables have been deleted. The tree t^* has thus at most $L(f) - 1$ leaves and computes f , which is impossible since $L(f)$ is the complexity of f .

Let us denote by t_e the subtree rooted at v the lowest common ancestor of the wildcards. We have shown that a typical tree computing f is a minimal tree of f in which we have plugged an expansion tree t_e which does not change the function f . \square

Lemma 13. *Let t be a typical tree computing f . The expansion tree t_e is either a simple tautology (or simple contradiction), or an x -expansion - i.e. a tree with one \wedge -leaf (resp. \vee -leaf) labelled by an essential variable of t .*

Proof. As shown in the former lemma, a typical tree computing f is a minimal tree of f on which has been plugged an expansion tree t_e .

First Case: Let us assume that t_e has no $(N \oplus P)$ -repetition and no essential variable among its $(N \oplus P)$ -pattern leaves. Then, we can replace t_e by a wildcard and simplify this wildcard. This simplification suppresses at least one other leaf of the tree: the obtained tree is then smaller than the original minimal tree, and still computes f . It is impossible.

Second Case: Let us assume that t_e has at least two $(N \oplus P)^2$ -restrictions. Thanks to Lemma 10, this family of expanded trees is negligible.

Third Case: Let us assume that t_e has exactly one $(N \oplus P)^2$ -restrictions. Then it must be a $N \oplus P$ -restriction (cf. First Case).

- if it is a repetition, than one can show that it must be a simple tautology or a simple contradiction.
- if it is an essential variable, one can show that it must be an X -expansion.

\square